# A Framework for Automatic Personalised Ontology Learning

Md Abul Bashar, Yuefeng Li
*Electrical Engineering and Computer Science School*
*Queensland University of Technology (QUT)*
*Brisbane, QLD 4001, Australia*
*Email: {m1.bashar, y2.li}@qut.edu.au*

Yang Gao
*Beijing Institute of Technology*
*Haidian, Beijing, China*
*Email: gyang@bit.edu.cn*

***Abstract*—Understanding or acquiring a user's information needs from their local information repository (e.g. a set of example-documents that are relevant to user information needs) is important in many applications. However, acquiring the user's information needs from the local information repository is very challenging. Personalised ontology is emerging as a powerful tool to acquire the information needs of users. However, its manual or semi-automatic construction is expensive and time-consuming. To address this problem, this paper proposes a model to automatically learn personalised ontology by labelling topic models with concepts, where the topic models are discovered from a user's local information repository. The proposed model is evaluated by comparing against ten baseline models on the standard dataset RCV1 and a large ontology LCSH. The results show that the model is effective and its performance is significantly improved.**

***Index Terms*—Ontology Mining, Personalisation, User Information Needs, Labelling Topic Models, Web Intelligence.**

## 1. Introduction

Over the last few decades, the amount of information available on the Web has increased exponentially. As a result, gathering useful information from the Web has become challenging. To Make matters worse, traditional search engines return the same search results to different users for the same query [1]. Two users may not have the same interests and preferences even though they use the same query. For example–for a query *java*, two users who are searching for *programming-language* and *coffee*, respectively, should not get the same result. Different and context focused results should be returned for each user [1]. To facilitate this, Web information gathering systems have to determine each user's information needs so that they can provide the right information tailored to specific users [2], [3]. Most of the time users cannot explicitly describe their information needs [2], so the system must have methods of doing so in the background. In this context, and for many other applications, understanding or acquiring a user's information needs from their *local information repository* (a set of example-documents that are relevant to the user's information needs) is important. However, acquiring a user's information needs from their local information repository is very challenging.

To discover a user's information needs from their local information repository, three approaches have been proposed in current literature–(a) bag-of-words [1], [4], [5], [6], (b) frequent text patterns [3], [7] and (c) topic models [8], [9]. The bag-of-words approach is simple but provides poor representation of a user's local information repository [3]. It cannot preserve the associations of terms in documents, and therefore cannot capture the user's intentions effectively. To address this problem, Li et al. [3] proposed to use frequent text patterns discovered from users' local information repositories. However, pattern mining produces large number of patterns, and using them effectively is difficult. Wu et al. [7] and Li et al. [3] made a break through in utilising patterns by deploying them to a term space. However, deploying patterns to a term space ultimately leads to bag-of-words. Other researchers [8], [9] use topic models to capture each user's information needs. Topic modelling is one of the most popular approaches for inferring the subject matter of a collection of documents [10], [11]. It discovers the statistical structure that corresponds to semantic themes present in the collection [9]. Some researchers argue that it has the ability to capture user interests [8], [9], and it can cluster groups of co-occurring terms [10]. Others suggest that the topic modelling approach is promising for search engines [10]. However, most of the discovered topic models do not produce easy-to-understand semantic meanings [11], [12], [13]; as a result, acquiring a user's information needs is still far from ideal.

Web ontologists observed that users implicitly possess some conceptual-models when they (the users) are gathering information from the Web [3]. The conceptual-models guide them to decide whether a document is relevant to them. Ontologies are considered powerful tools for simulating the conceptual models [1] because of their expressiveness, effective knowledge representation formalism and associated inference mechanisms. An ontology consists of a set of concepts and their semantic relations (e.g. Is-a, Related-to, Part-of), where a concept is a non-empty set of terms that together express a human-understandable meaning. Based on these observations, Tao et al. [2] at QUT (Queensland University of Technology) proposed a framework [2] to learn a personalised ontology semi-automatically, where a standard ontology is used as a source of concepts and their

semantic relations. The ontology is personalised based on the user's interaction with the ontology and meta data in their local information repository, and therefore it can effective simulate the conceptual model of user's information needs [2].

However, even though the framework proposed in [2] is an effective way to acquire user information needs, but it is only semi-automatic and requires meta data. Many local information repositories do not have meta data, and semi-automatic construction of personalised ontology is expensive and time-consuming. The open research question is how to automatically learn an effective personalised ontology from a user's local information repository that does not have meta data. In the given framework, the crucial part of learning the personalised ontology is selecting a set of concepts (and the semantic relations between the concepts) that can capture a user's information needs. Mapping the documents available in the local information repository to the standard ontology results in a huge amount of concepts, which causes the performance of the system decline dramatically (see section 5 for the performance of such a model named LDA-based-concept).

To address the research question, this paper proposes a model, TLPO (Topic-Model Labelling based Personalised Ontology), to automatically learn a personalised ontology. It learns a user's personalised ontology in three steps–(a) discovers a set of topic models from the user's local information repository, (b) labels the topic models with concepts and (c) uses the concepts and their semantic relations to construct a personalised ontology.

The leading idea of labelling topic models is to map them to the concepts in a standard ontology. The set of concepts that are mapped with topic models is selected as the set of labels. Existing mapping techniques (e.g. [1], [4], [5], [6]) can map only an individual term to concepts. That means, if we use existing techniques, instead of a topic model itself, the terms of the topic model are mapped to the concepts. Therefore, the associations of terms in the topic model are not reflected in the labelling, while the terms' associations are one of the most important features of the topic model. More importantly, each term of a topic model is mapped to a set of concepts, but working out how to use these mapped concepts to label the topic model is another challenge. The union ($\cup$) of the sets of concepts will result in too many concepts, while many of them are noisy, which means there are too many irrelevant concepts. On the other hand, the intersection ($\cap$) will result in too few or no concepts. Using the combination of union and intersection will present the system with too many combinations to consider, and no priorities on which combinations to be considered. As a result, selecting concepts for labelling topic models is challenging. To address this challenge, we propose an algorithm that can map the topic model itself, and therefore reflects the term association of the topic model in the labelling. To remove noisy concepts, it categories candidate concepts into groups based on their likelihood of noise.

This paper makes two major contributions: (a) leverages the local information repository to identify relevant concepts and (b) proposes a new effective framework for personalised ontology learning.

A set of experiments in information filtering using the model TLPO on a standard dataset RCV1 and a large ontology LCSH is conducted in this research. The experiments demonstrate the comparison between 10 baseline models and the proposed TLPO model. The evaluation results prove that the TLPO model can successfully learn personalised-ontologies by labelling topic models. It achieves significant improvement compared to the baseline models.

## 2. Related Works

Sometimes, topic models are manually labelled with concepts [14], [15], but the manual approach is expensive, time consuming and subjective. For automatic labelling, Mei et al. [12] and Hulpus et al. [10] propose to label topic models in terms of n-grams and phrases that are extracted from a corpus. They assume that extracted n-grams and phrases are semantically meaningful, therefore should express the topic themes. However, in many experiments, it has been observed that automatically extracted phrases and n-grams are not semantically meaningful especially for more than 2-grams [12]. Lau et al. [16] propose to find the best term of a topic and use it as the label. Single terms are too general, and therefore cannot accurately capture the themes of a topic [12], [17].

Hulpus et al. [10] and Lau et al. [17] propose to map topic models to concepts for semantic meaning. Other researchers (e.g. [11], [13], [18], [19], [20]) propose to map documents to concepts. The proposed techniques of these researchers potentially map a huge number of concepts, where many of them are irrelevant or noisy.

For example, in the ESA (Explicit Semantic Analysis) technique [19], [20], a text fragment is represented as a vector of concepts, where a concept is mapped based on the similarity between the text fragment and the concept-gloss (content of the Wikipedia article). The similar technique is used by [17] for labelling topic models. However, Egozi et al. [21] pointed out that the quality of concepts generated by ESA was lower than expected. Egozi et al. [22] identified that while some of the mapped concepts are relevant but many of them are not. Several incidental mentions of a term (from the text fragment) in the gloss is sufficient to trigger these noisy concepts. This problem will be severe for very small term sets like top terms in topic models. Such noisy concepts will lead to an interpretation that is completely wrong to the intention of the topic model. Also, many ontologies (e.g. LCSH (Library of Congress Subject Headings)) do not have glosses.

Chemudugunta et al. [11] and [13] use probabilistic methods and Gabrilovich et al. [18] use text categorisation technique for mapping documents to concepts while Hulpus et al. [10] use eigenvalue based measures for mapping topic models to concepts. All the existing techniques consider all the mapped concepts as useful, but some of the concepts are noisy. They use numerical scores to rank and select top concepts, but they do not have any mechanism to clearly differentiate between noisy concepts and noise-free concepts.

A mechanism addressing this problem will be significant contribution in topic model labelling.

The gloss problem of ESA was addressed to some extent in the model POM (Personalised Ontology Model) [5]. However, it assumes that terms are independent in a text fragment, which is not the case. Terms have associations between them. As a result, the mapped concepts are not effective (the experimental result of this model is shown in section 5.6).

## 3. Basic Definitions

The proposed model TLPO and the corresponding experiments in this paper are designed from the perspective of user information needs. In this section we give the basic definitions that are important for understanding the proposed model TLPO.

Let a user provide a set of example-documents that are relevant to their information needs. Because these documents are explicitly selected by the user as relevant to the subject matter of their interest, we use this document set as the source to capture the user's information needs. These documents constitute the user's local information repository. We use topic modelling as a tool to reduce the dimensionality of the documents and to discover the hidden but useful user information needs that the user cannot express explicitly. The output of the topic modelling is used to select a set of concepts and their semantic relations from a standard ontology. That is, a set of concepts that corresponds to the user's personal interest is selected. After that, the selected concepts and their semantic relations are used in constructing the personalised ontology.

The standard dataset RCV1 (Reuters Corpus Volume I) of TREC-10/2001 filtering track [7], [23], [24] is used in this research. It has a number of topics, and corresponding documents, we call these TREC-topics to avoid confusion with topics in topic modelling. Besides, each TREC-topic has a manual specification of information needs written by linguists. For each TREC-topic, domain experts divided the documents in dataset into a training set and a testing set. They further divided each of the training sets and testing sets into positive and negative sets. The positive set consists of documents that are relevant to a TREC-topic specification, and the negative set consists of documents that are not relevant to the TREC-topic specification (more details in section 5.1).

The set $D^+$ of positive documents in the training set that are relevant to a TREC-topic is used as an example-document set (this is to incorporate the real life fact that people usually do not provide negative documents). We extract topic models from these example-documents.

In this research, we choose the popular topic modelling technique Latent Dirichlet Allocation (LDA) which is described in the following subsection.

### 3.1. Latent Dirichlet Allocation

Let $D^+ = \{d_1, d_2, ..., d_M\}$ be a collection of $M$ relevant documents that constitutes the local information repository. In LDA, each document is considered as a bag-of-terms [11], [25]. Let $D_t = \{t_1, t_2, ..., t_V\}$ be the set of unique terms in the document collection $D^+$, where $V$ is the size of the vocabulary.

The idea behind LDA is that observed terms in each document are generated by a document-specific mixture of corpus-wide hidden topics [26]. It is a low-dimensional representation of documents. The number of hidden topics are assumed to be fixed to $T$, where this research uses $T$ = 10. A topic $z_j$ is represented as a multinomial probability distribution over the $V$ terms as $p(t_i|z_j)$, where $1 \leq j \leq T$ and $\sum_i^V p(t_i|z_j) = 1$. A document $d$ is represented as probabilistic mixture of topics as $p(z_j|d)$. Therefore, the probability distribution of $i$th term in a document $d$ can be model as a mixer over topics: $p(t_i|d) = \sum_{j=1}^{T} p(t_i|z_j)p(z_j|d)$. Here the only observable variable is $p(t_i|d)$. The other two variables $p(t_i|z_j)$ and $p(z_j|d)$ are hidden. In this paper, the widely used [27] statistical estimation technique of Gibbs sampling is used for learning the hidden variables. For more details on LDA, interested readers are referred to [11], [25], [26].

Usually people use top terms for representing a topic [12], [16], [17]. In most cases top 10 terms are sufficiently representative of the whole set of terms in a topic [16]. Therefore, in this paper, we represent a topic with top 10 terms, ranked by the multinomial distribution $p(t|z)$. From now on, we refer to the top tern terms when we refer to a topic.

However, even though the topic modelling (LDA) has the potential to be used for learning user interests, it lacks semantic focus [11] and a global view [12], [13]. As it does not focus on the semantics, it can capture the essence of a document only to a limited extent. While a document is expressed assuming prior knowledge, topic modelling assumes that a document is what it has [13]. The semantic theme discovered by topic modelling can better be represented and understood in terms of concepts [17]. In the following subsection we discuss and define the concept.

### 3.2. Concept

Ideally, a concept is defined by a set of attributes, and it represents an abstract class of ideas or objects. Chemudugunta et al. [13] identify the concept as a non-empty set of terms that together express a human understandable meaning. Humans use their knowledge and judgement to manually select the terms in a concept based on semantic similarity [13] so that together they can represent a meaning. A concept can represent semantically rich notions [13], and it is interpretable, broader in coverage [12], [13] and has a global view. Also, the concepts can serve humans to organise and share their knowledge [19]. Based on the characteristics of topic models and concepts, Chemudugunta et al. [13] argue that there are natural relations between topic models and concepts [13]. It follows that labelling topic models with concepts provides a bridge for learning personalised-ontology.

In this research, concepts are selected from a standard ontology, where the standard ontology consists of a set of concepts and a set of semantic relations between the concepts. Three semantic relations are considered in this

research, they are: 'Is-a', 'Related-to' and 'Part-of'.

**Definition 1** (Standard Ontology)**.** *A standard ontology (or simply an ontology) is a pair $\langle E, R\rangle$, where $E$ is a finite set of concepts, and $R$ is a set of triplet $\langle c_1, c_2, r\rangle$, where $c_1$ and $c_2$ are two concepts and $r$ is their semantic relation.*

We use a large knowledge base LCSH [28] as the standard ontology. The LCSH classification comprises a thesaurus of subject-headings covering one of the most exhaustive topic lists in the world, and specifies the semantic relations between the subject-headings in the taxonomy. Comparing to other subject classification/categorisation systems, such as Dewey Decimal Classification (DDC), and Reference Categorisation (RC), the LCSH classification has superior features.

The LCSH classification has more subject classes (LCSH has 394,070, DDC has over 1000 and RC has over 100,000 subjects), a more complex structure (LCSH has a depth of 37, DDC has a depth of 3 and RC has a depth of over 10), and more detailed semantic relations (LCSH has Is-a, Related-to, Part-of; DDC has Is-a; and RC has Is-a) specified. These features make the LCSH a great description of knowledge and ontology backbone.

The subject-headings in LCSH are explicitly defined by domain experts, and therefore they are easily understandable by humans. That is, subject-headings in LCSH correspond to the concepts identified by Chemudugunta et al. [13]. Using LCSH as the standard ontology, a concept is formally defined as the following:

**Definition 2** (Concept)**.** *A concept $c$ consists of a set of attributes and represents an abstract class of ideas or objects. Each concept is labelled with a subject-heading $s$ from LCSH, where $s = \{t_1, t_2, ..., t_n\}$ is a set of terms. Each term $t \in s$ represents an attribute of the concept c, and the label is referred as $label(c) = s$.*

From now on, we use the terminology 'concept' and 'subject-heading' interchangeably. When we say 'attribute of a concept' or 'term of a concept', we mean 'term in the label of a concept' i.e. by $t \in c$ we mean $t \in s$.

## 4. Proposed Model

The proposed model, TLPO, can be summarised as follows–(1) a set $Z$ of topic models is extracted from the local information repository $D^+$, (2) each topic is represented by top ten terms ranked by probability distribution, (3) the topic set $Z$ is labelled with a set of concepts (see section 4.1), (4) a set of smallest-upper-bound concepts (see section 4.2) are extracted from the standard ontology, (5) semantic relations of the concepts (both the labels and the smallest-upper-bound concepts) are extracted from the standard ontology and (6) the personalised-ontology is constructed from the concepts and their semantic-relations using definition 5. In the following subsections, we give a detailed description of the proposed model.

### 4.1. Labelling Topic Models

In general, there is a many-to-many relation between the concepts and the topics–a topic may be related to many concepts, and a concept may be related to many topics. Therefore, selecting a set of concepts as the set of labels is difficult [29]. As we discussed in section 1, existing mapping techniques (e.g. [1], [4], [5], [6]) cannot use the associations of terms that exist in a topic model for label selection because they map each term $t$ in the topic model individually to the concepts, rather than mapping the topic model itself. To address the mapping question, we propose an algorithm that can map the topic model itself, and therefore reflects the association of terms in a topic model. The algorithm is based on the function in Equation 1.

The function in Equation 1 estimates and assigns a relevance score for each concept $c$. A concept that is assigned a relevance score of 1 is called a *exactly matched* concept, and a concept that is assigned a relevance score less than 1 but greater than 0 is called a *partially matched* concept. An exactly matched concept can represent the knowledge of topics precisely but a partially matched concept may incorporate some noise. That is, partially matched concepts can sometimes be irrelevant. The advantage of equation 1 is that it can measure a concept's relevance associated with the whole topic set $Z$ rather than an individual topic or an individual term in a topic. How much a concept is irrelevant to the topic set can be estimated using the equation $irrel(s) = 1 - rel(s)$.

$$\begin{aligned} &rel(s) = \tfrac{|s \cap z_i|}{|s|} \\ &where,\ z_i \in \arg\max_{z \in Z}(|s \cap z|) \end{aligned} \tag{1}$$

The relevance score estimated by Equation 1 is used in the Algorithm 1 for mapping a set of topics to a set of concepts. Firstly, we find all the subject-headings of LCSH, where the relevance score of $s$ is greater than 0; we call this set $S'$, the candidate concept set. The set $S'$ contains both the exactly matched and the partially matched subject-headings. Secondly, from $S'$, we select all the subject-headings with a relevance score equal to 1 (i.e. exactly matched concepts) and call this set $C^e$. Remaining subject-headings in $S'$ are the partially matched subject-headings (i.e. partially matched concepts). Finally, our goal is to select not less than $k$ top-relevant subject-headings, where $k = |Z| \times \theta$ and $\theta$ is an experimental coefficient. If the number of subject-headings selected is greater than or equal to $k$, then we are done. If the number of selected subject-headings is less than $k$, then the remainder of the $k$ subject-headings are selected from the partially matched subject-headings. Where the value for the remaining $k$ is $k' = k - |C^e|$. To reduce noise, all the partially matched subject-headings where $rel(s) \leq irrel(s)$ are discarded. Then, based on their relevance value, the top $k'$ of partially matched subject-headings are selected and call this set $C^p$. The set $C = C^e \cup C^p$ of concepts is the set of labels for the given set of topics.

There are some terms in the topics that do not match with any concepts in the ontology. We assume that the terms that do not match with any concepts in the ontology are new concepts, created by the author of the document. This assumption is supported by [30]. They argue that when a new term is introduced, it creates a new concept that is associated with a specific area of knowledge. These new concepts contain important knowledge of the document

**Algorithm 1** Concept Mapping Algorithm

**Input:**
A set $S$ of all the subject-heading in LCSH; a set $Z$ of topic models; experimental coefficient $\theta$.
**Output:**
A set $C$ of concepts relevant to $Z$.
Let $C = C^e = C^p = S' = S'' = \emptyset$;
For each $s \in S\{$
  IF($rel(s) > 0$) then
    $S' = S' \cup \{s\};\}$
Let $k = |Z| \times \theta$;
For each $s \in S'\{$
  IF ($rel(s) == 1$) then{
    $C^e = C^e \cup \{s\};\}$ }
IF ($|C^e| \geqslant k$) then { $C = C^e;\}$
Else {
  $S' = S' - C^e$; // partially matched concepts
  For each $s \in S'\{$
    IF($rel(s) > irrel(s)$) then {
      $S'' = S'' \cup \{s\};\}\}$
  Sort $S''$ in descending order using $rel(s)$ value;
  Let $k' = k - |C^e|$;
  $C^p = topConcepts(k', S'')$;
  $C = C^e \cup C^p;\}$
Return;

(observed in the experiments); therefore, they are added to the set of labels.

## 4.2. Personalised Ontology Learning

After labelling of topics is done (i.e. mapping topics to concepts), the next obvious question is how to represent the user's information needs using these concepts? The answer to this question is personalised ontology.

Use of ontology as a formal model for simulating the conceptual model of user's information needs appeared to be promising in researches done by [2], [3]. An ontology that is learned from a user's local information repository and captures personal preferences and interests is called a personalised ontology [2].

In this paper, we propose a new framework for personalised ontology. The framework has two structures: a Semantic Structure ($\mathcal{SS}$) and a Contextual Structure ($\mathcal{CS}$). The Semantic Structure defines the core of the personalised ontology. It includes concepts and their semantic relations. It is formally defined in Definition 3. On the other hand, Contextual Structure defines the context of a user's information needs, where the knowledge of the user's local information repository and the given standard ontology is combined. The Contextual Structure of the framework is formally defined in Definition 4. Using these two structures, the personalised ontology is defined in Definition 5.

**Semantic Structure:** Two concepts are called semantically related if they have a semantic relation such as 'Is-a', 'Related-to', 'Part-of', etc. For example, if $c_1$ and $c_2$ are two concepts and $c_1$ Is-a $c_2$ (or vice versa), we say that they are semantically related. An implied-semantic relation means: either the semantic relation explicitly exists in an ontology or it can be inferred from the semantic relations in the ontology (for example: if 'cat Is-a mammal' and 'mammal Is-a vertebrata' are two semantic relations in the ontology, then we can infer that 'cat Is-a vertebrata').

**Definition 3** (Semantic Structure)**.** *A Semantic Structure is a triplet $\langle MC, SC, \mathbb{R}\rangle$, where $MC$ is a set of concepts that are selected as labels for a set of topic models; $SC$ is a set of smallest upper-bound concepts that has implied-semantic relation with more than one $mc \in MC$; and $\mathbb{R}$ is a set of triplet $\langle c_1, c_2, \varphi\rangle$, where $c_1$ and $c_2$ are any two concepts in $MC \cup SC$ (such that $c_1 \neq c_2$), and $\varphi$ is a semantic relation between $c_1$ and $c_2$.*

Let in the ontology, $c'$ and $c''$ be two concepts that have implied-semantic relation with more than one $mc$, $L_1$ be a non empty set of $mc$ that has implied-semantic relation to $c'$, $L_2$ be a non empty set of $mc$ that has implied-semantic relation to $c''$; $c'$ is called the smallest upper-bound concept if it is not an ancestor of $c''$, and $c''$ is called the smallest upper-bound concept if it is not an ancestor of $c'$, or each of $c'$ and $c''$ are called the smallest upper-bound concepts if $L_1 \neq L_2$.

**Contextual Structure:** Regarding the category of concepts (exactly matched and partially matched) that a term can appear in, we have three cases. Formally we can write the three cases as: $case_1$ = $(\exists c_1 \in C^e \ \& \ \exists c_2 \in C^p) \Rightarrow (t \in c_1 \cap c_2)$; $case_2$ = $(\exists c \in C^e \Rightarrow t \in c)$ & $(\forall c \in C^p \Rightarrow t \notin c)$; $case_3$ = $(\exists c \in C^p \Rightarrow t \in c)$ & $(\forall c \in C^e \Rightarrow t \notin c)$.

The contextual structure has five information levels. Following are the brief descriptions of the levels:

**Document Level Information:** Term frequency is related to the distribution of a term in the documents of a corpus, and therefore it is document-level information of the term. It indicates how important the term is in relation to the subject matter of a document set [31] thereby to the user preferences. Term frequency is the number of times a term, $t$, occurs in all the positive documents $D^+$, i.e. $f(t) = \sum_{d \in D^+} f(t, d)$. After normalising $f(t)$ by the total number of terms in all the documents $D^+$, we get a normalised term frequency, i.e. $f_r(t) = \frac{f(t)}{\sum_{d \in D^+} |terms(d)|}$, where $terms(d)$ returns all the terms in the document $d$. The $f_r$ implicitly utilises the structure of the documents.

**Topic Level Information:** In LDA, a document, $d$, in a user's local information repository is represented by a probabilistic mixture of topics as $p(z_j|d)$ [11], [25]. This probabilistic mixture can represent a user's interest in the topic. The full semantic theme of a topic $z_j$ is represented by its corresponding multinomial distribution over terms as $p(t_i|z_j)$ [12]. It can be assumed that a concept containing the high probability terms as its attributes is more closely associated to the topic theme [12], [17], [32]. Therefore, for a user, the amount of topical interest that an attribute contains can roughly be estimated as $w_z(t) = \sum_{j=1}^{T} p(z_j|d) \times p(t|z_j)$. This estimation is for a single document. In case of multiple documents (i.e. $D^+$), we take the average. The $w_z(t)$ implicitly utilises the structure of the topics.

**Inter-Topic Level Information:** From the experimental results, Mao et al. [32] concluded that inter-topic relations are useful for improving the accuracy of topic interpretation. To utilise the term overlapping between topics, the set $Z$ of topics is deployed on term space $T$ [7]. A deployment weight $w_\partial(t) = \frac{|\{z|t \in z, z \in Z\}|}{|Z|}$ can be calculated for each term in the term space. This weight is inter-topic level information, and it implicitly utilises the parent-child structure of topics [32].

**Ontology Level Information:** If a term appears in many

concepts in the standard ontology, the term is general. The specificity of a term is inversely related to the frequency of concepts in the standard ontology that contains this term [33]. On the other hand, the frequency of exactly matched concepts in the personalised ontology that contain a given term indicates how closely the term is related to the main theme of the personalised ontology. The ontological significance $spe_o(t) = \frac{|\{c|t \in c, c \in C^e\}|}{|\{c|t \in c,\ c \in LCSH\}|}$ of a term is estimated using these two frequencies. The $spe_o$ implicitly utilise the structure of the ontologies.

**Mapping Level Information:** The covering set for $c$ is the set of all the topics $z \in Z$ such that $c \cap z \neq \emptyset$. That is $coverset'(c) = \{z | z \in Z, c \cap z \neq \emptyset\}$. The *support* for the concept $c$ is $sup'(c) = \sum_{z \in coverset'(c)} \frac{|c \cap z|}{|c|}$. Concept support indicates how closely a concept and the topic set is related. Based on this concept support, the overall relatedness of a term to both the topic set and the concept set can be estimated using the following equation of $i(t)$.

$$i(t) = \begin{cases} i_1(t) & if\ case_2 \\ i_2(t) & if\ case_3 \\ \frac{i_1(t)+i_2(t)}{2} & if\ case_1 \end{cases}$$

$$i_1(t) = \frac{\sum_{t \in c, c \in C^e} \left\{ \frac{sup'(c)}{|c|} \right\}}{|\{c \in C^e | t \in c\}|}$$

$$i_2(t) = \frac{\sum_{t \in c, c \in C^p} \left\{ \frac{sup'(c)}{|c|} \right\}}{|\{c \in C^p | t \in c\}|}$$

**Definition 4** (Contextual Structure)**.** *A Contextual Structure is a tuple $\langle DLI, TLI, ILI, OLI, MLI \rangle$, where $DLI$ is Document Level Information, $TLI$ is Topic Level Information, $ILI$ is Inter-topic Level Information, $OLI$ is Ontology Level Information and $MLI$ is Mapping Level Information.*

**Personalised Ontology:** To better understand user information needs, the personalised ontology is defined in terms of both the Semantic Structure and the Contextual Structure. It helps us to know both the conceptual model and the context that shapes the conceptualisation.

**Definition 5** (Personalised Ontology)**.** *A Personalised Ontology is a pair $\langle \mathcal{SS}, \mathcal{CS} \rangle$, where $\mathcal{SS}$ is a Semantic Structure and $\mathcal{CS}$ is a Contextual Structure.*

# 5. Evaluation

The hypothesis of this research is that *a personalised ontology that can effectively acquire user information needs can be learned by labelling a set of topic models with concepts, where the set of topic models is extracted from a user's local information repository*. However, it is difficult to evaluate the quality of the learned personalised ontology [2], [34]. One possible way is manual checking [34] (e.g, manually checking whether the ontology or part of it represents user information needs [34]). Unfortunately, manual checking is subjective and very expensive, even impossible for a large dataset [34].

Because of inherent difficulties of evaluating the effectiveness of an ontology, Brewster et al. [34] propose to decompose the ontology into its constituent parts. In the simplest form, an ontology consists of a set of concepts and their relations. Bloehdorn et al. [35] proposed to use the concepts in an ontology for text classification as a way to evaluate a learned ontology. Brewster argue that the constructs of an ontology can be viewed as the abstractions of natural language texts. For evaluation, they propose to revise the abstraction by finding the signatures of these constituents in the natural language texts. Inspired by these works, we propose to find the signatures of the concepts (the constituent parts of semantic structure) and the contextual structure in unknown documents to check the unknown documents' relevance to user information needs.

Based on the contextual structure analysis, we estimate a single weight for each term in the concepts. The weight can be viewed as a quantitative digest of the contextual structure. The main goal of this weighting is to utilise the essential statistical relationships that exist in the contextual structure. We estimate the term weight using the following Equation 2.

$$w(t) = \begin{cases} w_z(t) \times \alpha_1 + \beta_1 \times i(t) \times w_c(t) & if\ case_1 \\ w_z(t) \times \alpha_2 + \beta_2 \times i(t) \times w_c(t) & if\ case_2 \\ w_z(t) & if\ case_3 \end{cases} \quad (2)$$

$$w_c(t) = \begin{cases} spe_o(t) \times f_r(t) & if\ case_1 \\ spe_o(t) \times f_r(t) \times w_\partial(t) & if\ case_2. \end{cases}$$

Here, $\alpha_1$, $\alpha_2$, $\beta_1$, and $\beta_2$ are experimental coefficients. We estimate the term weight of new concepts (see section 4.1) using $w(t) = w_z(t)$.

The contextual structure of the personalised ontology is inherently represented by the assigned term weight. Therefore, to prove the hypothesis, we need to show that the concepts in personalised ontology and the assigned term weight are effective for the information gathering system.

## 5.1. Data Collection

The standard dataset RCV1 of TREC-10/2001 Filtering Track [7], [23], [24] and the large ontology LCSH are used in the research experiments. RCV1 consists of 806,791 news stories (one story per document) provided by Reuters, LTD [7], [24]. It has 100 TREC-topics, where each TREC-topic contains different numbers of documents. The documents in the first 50 TREC-topics are manually categorised by domain experts, and [36] argue that the first 50 TREC-topics are stable and sufficient for maintaining the accuracy of the evaluation measures. Therefore, the first 50 TREC-topics are used in this research.

The 'story title' and 'story text' are used as the content of one document. Pre-processing is applied to all the documents via meta-data and stop-words removing as well as stemming. We use only the positive documents of the training set for extracting topic models and training other baseline models, while both the positive and the negative documents in the testing set are used for evaluation.

## 5.2. Baseline Models

In order to provide a comprehensive evaluation of our proposed model, we have selected 10 baseline models in four different categories that are shown in Table 1.

## 5.3. Evaluation Measures

Our proposed model TLPO is evaluated by different means. Especially, five widely used measures of information filtering that are based on relevance judgements. They include the Mean Average Precision $(MAP)$, the average precision of top 20 returned documents $(T20)$, the $F_{score}$ measure $(F_1)$, the break-even point $(BP)$, and the interpolated precision averages at 11 standard recall levels $(11 - point)$.

TABLE 1: Baseline Models

**Category 1: Topic Modelling Based**

*LDA-word* [25], [26], [37]: uses the term frequency to represent topic relevance and the association of terms with different topics to represent user interests.

*TNG* [38]: is an $n$-Gram based topic model.

**Category 2: Concept Based**

*LDA-based-concept* [11]: uses statistical LDA technique for annotating text documents with the concepts. It treats concepts as topics with constraint $w_i \notin c_j \Rightarrow p(w_i|c_j) = 0$.

*POM* [5]: is one of the most recent work that maps document keywords to the standard ontology (LCSH).

**Category 3: Pattern Based**

*Pattern Deploying Model (PDM)* [7]: provides a way to effectively use the text-patterns in the information filtering.

*FCP* [26]: frequent closed patterns extracted from documents are used to represent user interests.

*Master Pattern (MP)* [39]: is a profile-based technique for summarising a collection of frequent closed patterns, using only $K$ representatives. It is popularly used in data mining communities for effective utilisation of patterns.

*n-Gram* [26]: uses $n$-Grams extracted from documents to represent user interests, where $n$ is empirically set to 3.

**Category 4: Term Based**

*BM25* [40]: is one of the term-based state-of-the-art models for representing documents.

*Support Vector Machine (SVM)* [41]: is considered effective for text filtering and categorisation.

## 5.4. Experimental Design

In the TREC Filtering Track [7], [23], [24], when testing a system, the user's information need is assumed stable and a stream of unknown documents (from the testing dataset) is brought into the system. For each new document, the system has to decide whether the document is relevant to the user's information needs [23].

As discussed in section 1, many Web ontologists observed that every user possesses implicit conceptual-models that guide them to judge whether a document is relevant to their information needs [2], [3]. Based on this observation, in this paper, we propose an objective evaluation approach where the personalised ontology is used as a conceptual-model. That means, a machine uses the personalised ontology to predict whether a new document brought into the system is relevant to the user's information needs. Brewster et al. [34] argue that a good ontology can serve its purpose, and Calegari et al. [1] argue that the more effectively a personalised ontology represents user information needs, the higher the probability to improve information gathering performance is. If the machine can predict the relevance, we believe it indicates that the personalised ontology can effectively represent the user's information needs. It is a data-driven evaluation of ontology in a real application as suggested by [34]. In the context of machine readability of the Web in the future, this king of evaluation is appropriate [34].

To prove the hypothesis, a series of experiments have been conducted on the standard dataset RCV1, using TREC-topics [23]. We use the mapped concepts and the assigned term weight as a query ($Q$) submitted to an information filtering system. A similar approach is applied for the baseline models. If the results of information filtering measures are significantly improved, compared with the baseline models, we can claim that our proposed model TLPO can learn personalised ontology that can acquire user information needs effectively.

## 5.5. Experimental Settings

In this paper, for all LDA-based topic models, the parameters are set as follows: the number of iterations of Gibbs sampling is 1000, the hyper-parameters of the LDA are $\alpha = 50/V$ and $\beta = 0.01$. These parameter values were used and justified in [42]. For extracting frequent closed patterns, the minimum support is sensitive to a given data set. For the RCV1 data set, using trial-and-error, the best value for this experimental coefficient was found to be 0.2. The best values for other experimental coefficients were also determined on the trial-and-error basis in RCV1. The best value for $\beta$ was found to be $0.2$ for generating master patterns; in the concept mapping algorithm, the best value for $\theta$ was found to be 3.0; and for weighting the terms of the concepts mapped by the TLPO, the best value for $\alpha_1$, $\beta_1$, $\alpha_2$ and $\beta_2$ were found to be 2.1, 55, 1.5 and 56, respectively, in the experiment.

## 5.6. Experimental Results

Evaluation results of the personalised ontology are shown in table 2 and figure 1. The results are the average of all the TREC-topics in the dataset. The table and figure also show the results of the 10 baseline models. The $change\%$ in table 2 means the percentage change of our proposed TLPO model over the best results of the baseline models. An improvement greater than $5\%$ is considered significant.

Table 2 shows that the information filtering performance of our proposed model TLPO is significantly better than the best results of the baseline models. It improved the performance significantly up to 5.839% (4.345% min and 11.180% max) in percentage change on average for all five measures. The amount of improvement is significant for all the individual measures too, except for the $F_1$ (in this case $4.345\%$ improved). The most important measure of information filtering is $MAP$. The model improved the $MAP$ performance significantly up to 6.629% in percentage change. The $11-point$ results in figure 1 show that the performance is consistently better than the baseline models.

TABLE 2: Evaluation Results

| | $Top-20$ | $BP$ | $MAP$ | $F_1$ |
|---|---|---|---|---|
| TLPO | 0.537 | 0.458 | 0.473 | 0.459 |
| LDA-word | 0.483 | 0.428 | 0.444 | 0.439 |
| PDM | 0.473 | 0.417 | 0.438 | 0.436 |
| POM | 0.458 | 0.400 | 0.411 | 0.419 |
| SVM | 0.447 | 0.409 | 0.408 | 0.421 |
| BM25 | 0.434 | 0.339 | 0.401 | 0.410 |
| MP | 0.426 | 0.392 | 0.393 | 0.409 |
| TNG | 0.446 | 0.367 | 0.374 | 0.388 |
| n-Gram | 0.401 | 0.342 | 0.361 | 0.386 |
| FCP | 0.428 | 0.346 | 0.361 | 0.385 |
| LDA-based-concept | 0.335 | 0.329 | 0.326 | 0.352 |
| **change%** | **11.180** | **7.043** | **6.629** | **4.345** |

A system is significantly different from another system if the $p$ value of tTest is less than 0.05 [43]. The one tailed $tTest$ results for the model TLPO compared with the best results of baseline models are given in table 3. Table 3 shows that, in all the measures, the $p$ values are less than 0.05. This implies that the performance improvement of the proposed TLPO model is statistically significant.

Based on these results, we can claim that our proposed

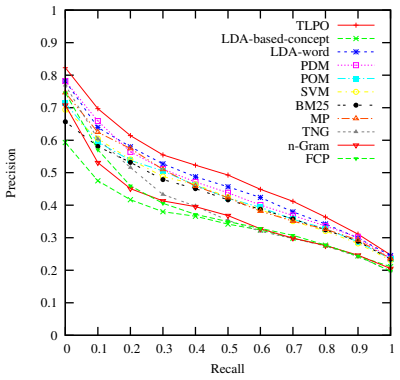

Figure 1: 11-Point Results

TABLE 3: tTest $p$-values

| $Top-20$ | $BP$ | $MAP$ | $F_1$ |
|---|---|---|---|
| 0.0435 | 0.0493 | 0.0412 | 0.0165 |

model TLPO learns a personalised ontology that can acquire user information needs effectively, which means the obtained results support our hypothesis.

## 6. Conclusions

This paper proposes a model, TLPO, for learning a personalised ontology that can acquire a user's information needs effectively. The model integrates information of both a standard ontology and a user's local information repository in a single conceptual model. As the personalised ontology is learned from a set of example documents that are relevant to the user's information needs, it will be useful for improving the performance of information gathering tailored to the specific user. The paper makes two major contributions–(a) leverages the local information repository to identify relevant concepts and (b) proposes a new effective framework for personalised ontology learning.

## Acknowledgments

This research was partially supported by Grant DP140103157 from the Australian Research Council (ARC Discovery Project).

## References

[1] S. Calegari and G. Pasi, "Personal ontologies: Generation of user profiles based on the yago ontology," *Information processing & management*, vol. 49, no. 3, pp. 640–658, 2013.

[2] X. Tao, Y. Li, and N. Zhong, "A personalized ontology model for web information gathering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 4, pp. 496–511, 2011.

[3] Y. Li and N. Zhong, "Mining ontology for automatically acquiring web user information needs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 554–568, 2006.

[4] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, *Ontology-based interpretation of keywords for semantic search*. Springer, 2007.

[5] Y. Shen, Y. Li, and Y. Xu, "Adopting relevance feature to learn personalized ontologies," in *AI 2012: Advances in Artificial Intelligence*, 2012, pp. 457–468.

[6] A. Sieg, B. Mobasher, and R. Burke, "Web search personalization with ontological user profiles," in *CIKM'07*, 2007.

[7] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in *ICDM'06*, 2006.

[8] M. Harvey, F. Crestani, and M. J. Carman, "Building user profiles from topic models for personalised search," in *CIKM'13*, 2013.

[9] D. Andrzejewski and D. Buttler, "Latent topic feedback for information retrieval," in *KDD'11*, 2011.

[10] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in *WSDM'13*, 2013.

[11] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers, *Modeling documents by combining semantic concepts with unsupervised statistical learning*. Springer, 2008.

[12] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *KDD'07*, 2007.

[13] C. Chemudugunta, P. Smyth, and M. Steyvers, "Combining concept hierarchies and statistical topic models," in *CIKM'08*, 2008.

[14] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *WWW'06*, 2006.

[15] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *KDD'06*, 2006.

[16] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, "Best topic word selection for topic labelling," in *COLING'10*, 2010.

[17] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2011, pp. 1536–1545.

[18] E. Gabrilovich and S. Markovitch, "Harnessing the expertise of 70, 000 human editors: Knowledge-based feature generation for text categorization," *Journal of Machine Learning Research*, vol. 8, no. 10, pp. 2297–2345, 2007.

[19] ——, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *IJCAI'07*, vol. 7, 2007, pp. 1606–1611.

[20] ——, "Wikipedia-based semantic interpretation for natural language processing," *Journal of Artificial Intelligence Research*, vol. 34, no. 2, p. 443, 2009.

[21] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," *ACM Transactions on Information Systems*, vol. 29, no. 2, p. 8, 2011.

[22] O. Egozi, E. Gabrilovich, and S. Markovitch, "Concept-based feature generation and selection for information retrieval," in *AAAI'08*, vol. 8, 2008, pp. 1132–1137.

[23] S. E. Robertson and I. Soboroff, "The trec 2002 filtering track report," in *TREC*, 2002.

[24] T. Rose, M. Stevenson, and M. Whitehead, "The reuters corpus volume 1–from yesterday's news to tomorrow's language resources," in *LREC'02*, vol. 2, 2002.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[26] Y. Gao, Y. Xu, and Y. Li, "Pattern-based topics for document modelling in information filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1629–1642, 2015.

[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[28] K. Yi and L. M. Chan, "Linking folksonomy to library of congress subject headings: an exploratory study," *Journal of Documentation*, vol. 65, no. 6, pp. 872–900, 2009.

[29] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, "Text mining and ontologies in biomedicine: making sense of raw text," *Briefings in bioinformatics*, vol. 6, no. 3, pp. 239–251, 2005.

[30] J. C. Sager, *A practical course in terminology processing*. John Benjamins Publishing, 1990.

[31] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *ICML'03*, 2003.

[32] X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, T.-S. Chua, H. Yan, and X. Li, "Automatic labeling hierarchical topics," in *CIKM'12*, 2012.

[33] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. A. Bijaksana, "Relevance feature discovery for text mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1656–1669, 2015.

[34] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, "Data driven ontology evaluation," 2004.

[35] S. Bloehdorn, P. Cimiano, and A. Hotho, "Learning ontologies to improve text clustering and classification," in *From Data and Information Analysis to Knowledge Engineering*. Springer, 2006, pp. 334–341.

[36] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *IR'00*, 2000.

[37] T. Hofmann, "Probabilistic latent semantic indexing," in *IR'99*, 1999.

[38] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in *ICDM'07*, 2007.

[39] X. Yan, H. Cheng, J. Han, and D. Xin, "Summarizing itemset patterns: a profile-based approach," in *KDD'05*, 2005.

[40] S. Robertson, H. Zaragoza, and M. Taylor, "Simple bm25 extension to multiple weighted fields," in *CIKM'04*, 2004.

[41] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[42] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[43] J. Wortsman, L. Y. Matsuoka, T. C. Chen, Z. Lu, and M. F. Holick, "Decreased bioavailability of vitamin d in obesity," *The American journal of clinical nutrition*, vol. 72, no. 3, pp. 690–693, 2000.